ИСПОЛЬЗОВАНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ DATA MINING ПРИМЕНИТЕЛЬНО К ПРОГНОЗИРОВАНИЮ ПАРАМЕТРОВ БЕЗОПАСНОСТИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

Ю.В. Передерин

В работе представлены результаты обработки базы данных по параметрам безопасности органических соединений классическими методами статистического анализа и при помощи нейронных сетей с приведением аргументов в пользу использования для указанных целей новых интеллектуальных систем программного обеспечения.

ВВЕДЕНИЕ

Современные технологии обрабатывают информацию с целью автоматического поиска шаблонов, характерных для каких-либо фрагментов неоднородных многомерных данных. В отличие от оперативной аналитической их обработки бремя формулировки гипотез и выявления необычных шаблонов переложено с человека на компьютер.

Специалисты на протяжении нескольких последних десятков лет решали подобные проблемы, но только сейчас общество в целом созрело для понимания практической важности и широты этих задач. Во-первых, в связи с развитием технологий записи и хранения данных сегодня на людей обрушились колоссальные потоки информации в самых различных областях, которые без продуктивной переработки грозят превратиться в ненужные свалки. И, во-вторых, средства и методы обработки данных стали доступными и удобными, а их результаты понятными любому человеку.

В настоящее время в практике для прогнозирования параметров безопасности взрывчатых веществ (чувствительность к различного рода воздействиям, температура вспышки) не существует однозначных методик, дающих приемлемые результаты по предсказанию значений этих параметров с удовлетворительной погрешностью – хотя бы менее 30%. Это связано с отсутствием математической модели, достоверно описывающей зависимость данных характеристик взрывчатых веществ (ВВ) от физико-химических свойств. Выявление математической модели для таких веществ можно проводить только с помощью метода кластеризации, что, в свою очередь, является довольно трудоемким процессом, имеющим приложение для узкого круга веществ.

По этой причине возрастает необходимость в системах, которые не только выполняют однажды запрограммированную последовательность действий над заранее определенными данными, но и способны сами анализировать вновь поступающую информацию, находить в ней закономерности, проводить прогнозирование и т.д.

МЕТОДЫ ИССЛЕДОВАНИЯ

К настоящему времени сложилось несколько продуктивных направлений развития методов индуктивного вывода, таких как:

– деревья решений;

– экспертные системы на базе нечеткой логики, задачи которых состоят в сборе, хранении и использовании знаний, полученных от экспертов, с целью решения прикладных задач идентификации. Оболочка состоит из двух основных частей: программной среды, позволяющей создавать экспертные системы в выбранной предметной области, и собственно экспертной системы, которая является конечным продуктом;

– кластерный анализ;

– нейронные сети.

Для каждого из этих направлений созданы функционирующие прикладные программные пакеты, имеющие узконаправленную специализацию, т.е. способные решать ограниченный круг практических проблем. Наибольший интерес для решения поставленных задач является использование нейронных сетей [1-5].

Цель создания нейросети – выявление скрытых правил и закономерностей в наборах данных. Дело в том, что человеческий разум сам по себе не приспособлен для восприятия больших массивов разнородной информации. Человек к тому же неспособен улавливать более двух-трех взаимосвязей даже в небольших выборках. Но и традиционная математическая статистика, долгое время претендовавшая на роль основного инструмента анализа данных, также нередко пасует при решении задач из реальной сложной жизни. Она оперирует усредненными характеристи-

ками выборки, которые часто являются фиктивными величинами (например, средняя температура пациентов по больнице). Поэтому методы математической статистики оказываются полезными, главным образом, для проверки заранее сформулированных гипотез.

В связи с широким применением (бизнес, медицина, химия, генная инженерия и т.д.) и наличием довольно удобных демоверсий программных пакетов, позволяющих делать прогнозирование параметров безопасности с довольно приемлемой погрешностью (1…2%), был выбран метод расчета характеристик веществ с помощью нейронных сетей, а именно при помощи программного пакета Deductor Studio.

Программа Deductor Studio предназначена для проведения исследований с целью выбора оптимальной конфигурации нейронной сети, позволяющей наилучшим образом решить поставленную задачу. Результатом работы системы является файл, который хранит в себе все параметры полученной нейронной сети. Далее, на основе этого файла, можно разрабатывать систему для решения конкретных задач. Для этого был разработан модуль, позволяющий работать с этим файлом.

В данный модуль включено несколько классов, предназначенных для создания нейросети, загрузки ее параметров из файла, созданного программой Deductor Studio и использования полученной нейросистемы. Скорость работы такого модуля полностью зависит от технических характеристик компьютера, на котором он установлен.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

При проведении работы была создана база данных свойств энергетических материалов с описанием ряда их свойств и показателей в качестве дескрипторов. За основу данных исследований было взято влияние атомарного состава молекулы $C_aH_bN_cO_d$ и ее молекулярной массы на чувствительность к удару (H_{50}). На первом этапе была проведена обработка существующей выборки методом матричных графиков в программном пакете Statistica 6.0 (рис. 1).

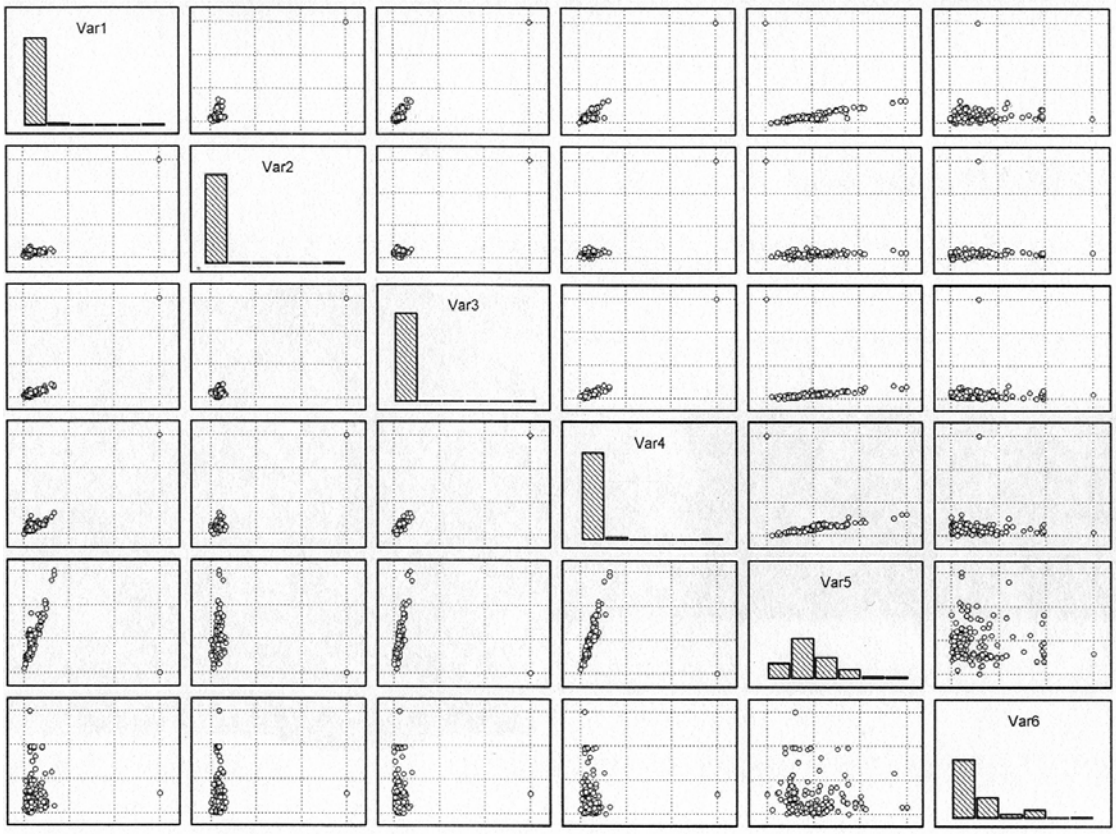


Рис. 1. Матричный график атомарного состава молекулы, молекулярной массы и чувствительности к удару (H_{50})

Из полученных данных видно, что обработка данных такого типа, когда аргументом является не один параметр, а более трех, с использованием линейной регрессии не дает

приемлемых результатов, а именно из полученных данных нельзя получить какую-либо полезную информацию.

При обработке существующей базы данных при помощи искусственной нейронной сети в программе Deductor Studio были получены результаты в следующем ниже графическом представлении (рис. 2).

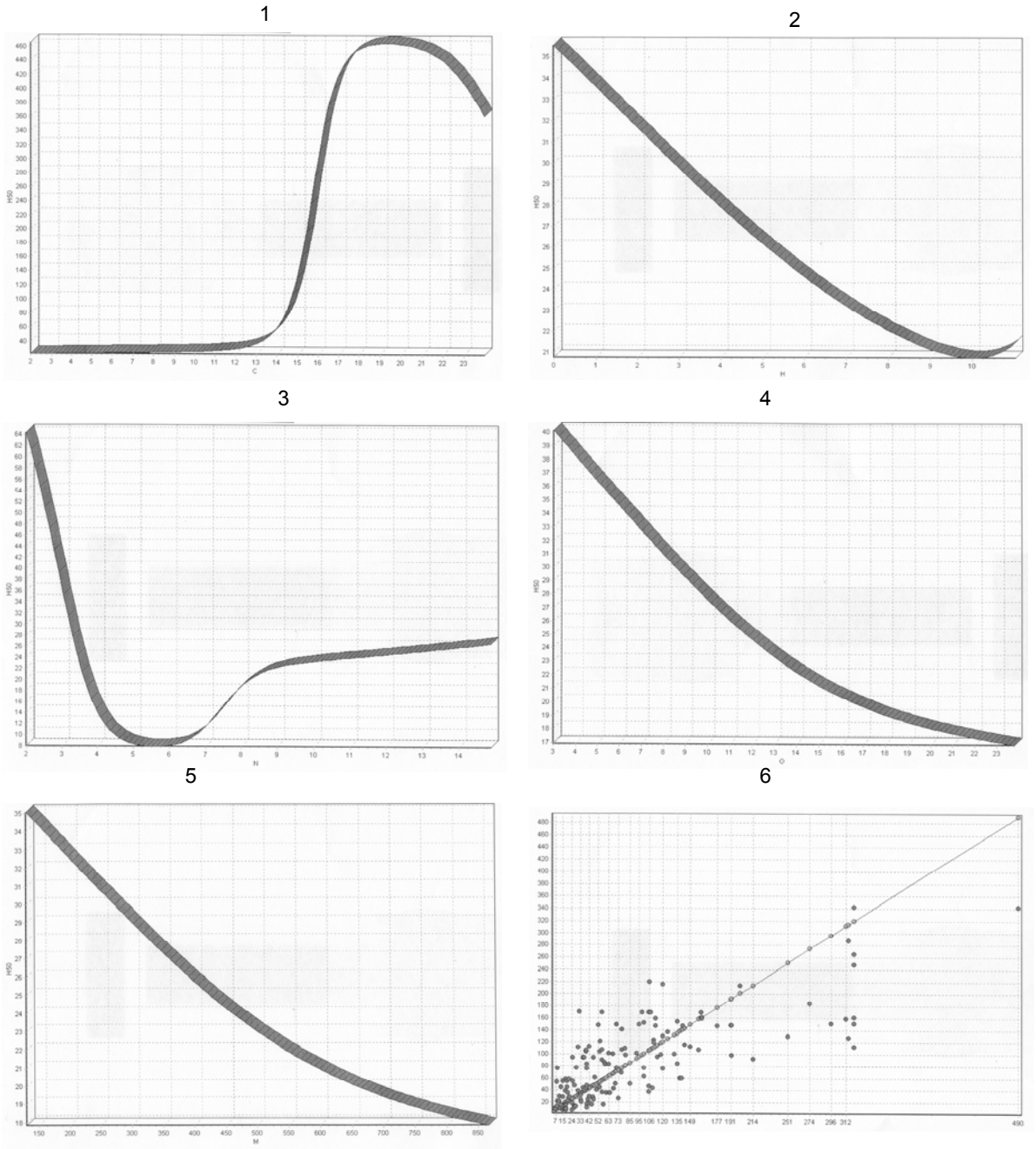
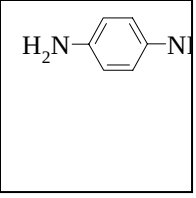


Рис. 2. Результат обработки имеющейся базы данных искусственной нейронной сетью – зависимости чувствительности к удару от содержания в молекуле: 1 – углерода; 2 – водорода; 3 – азота; 4 – кислорода; 5 – от молярной массы; 6 – диаграмма рассеяния прогнозируемых данных

На представленных зависимостях четко просматривается зависимость чувствительности энергетической структуры от выбранного параметра при постоянстве других пара-



метров, а диаграмма рассеяния показывает, что предсказанные значения чувствительности к удару в большинстве своем укладываются в полученную математическую модель, либо имеют допустимое отклонение.

На основании изложенного выше можно сделать вывод о том, что использование искусственных нейронных сетей является перспективным прикладным методом при обработке многомерных массивов данных в арсенале современных ученых и исследователей.

Предлагаемый подход позволяет получать конкретные значения выбранных параметров, основываясь на аппроксимации уже полученных данных в выбранной области с достоверностью, удовлетворяющей современным требованиям (1…4%).

ЗАКЛЮЧЕНИЕ

На основании полученных результатов можно с уверенностью говорить о том, что искусственные нейронные сети являются мощным инструментом в руках ученых и исследователей при обработке многомерных массивов экспериментальных данных и позволяют получать весьма ценную информацию там, где классические методы дают неудовлетворительные результаты.

ЛИТЕРАТУРА

1. Ротштейн А.П. "Интеллектуальные технологии идентификации" – Электронное издание.

2. Мишулина О.А., Манько С.В. Нейронные сети и устройства нечеткой логики // Приборы и системы. Управление, контроль, диагностика. – 2001. – №8. – С.36.

3. Гупал А.М., Понамарев А.А., Цветков А.М. Об одном методе индуктивного вывода с подрезанием деревьев решений // Кибернетика и системный анализ. –1993. – №5. – С.174.

4. Пожарная опасность веществ и материалов, применяемых в химической промышленности / Справочник. – М.: Химия, 1970. – 336 с.

5. Электронный учебник StatSoft – Электронное издание.